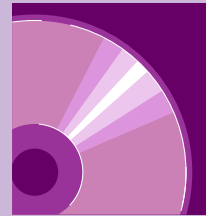


## ***SIPP Public Use Files***



*This section covers basic concepts and topics that analysts need to understand when working with the SIPP public use files.*

- *Types of SIPP Data Files*
- *Common Features Across SIPP Data Files*
  - Changes in Variable Names*
  - Survey Instrument Vs. Data Dictionary*
  - Identification/Description Variables*
    - Basic ID Variables*
    - Monthly Interview Status*
    - Identifying Persons*
    - Identifying Households*
    - Identifying Families*
    - Describing Relationships to Reference Persons*
    - Identifying Program Units*
    - Identifying Movers and Household Composition Changes*
    - Identifying States and Metro Areas*
  - Choosing Weights*
  - Income Topcoding*
  - Using Allocation Flags*

## Types of SIPP Data Files

There are three types of public use files containing SIPP data: core wave files, topical module files, and full panel longitudinal research files.

**Core Wave Files.** Since 1990, these files have been issued in person-month format. They contain up to four records for each primary sample member and for each person who ever lived with a primary sample member during the reference period. Each record contains data from 1 of the 4 reference months in the wave.



**Topical Module Files.** For the 1996 Panel, these files contain one record for each person who was in the sample with a completed or imputed interview in the fourth month of the wave's reference period. Topical module files from previous panels contain one record for each person who was in the sample with a completed or imputed interview during the interview month (month 5), not the fourth month of the reference period.

**Full Panel Longitudinal Research Files.** These files are also referred to as "full panel files" and "longitudinal files." They contain one record for each primary sample member and for each person who ever lived with a primary sample person during the panel.

## Common Features Across SIPP Data Files

The remainder of this section addresses features common to all three types of SIPP files. Although the features apply to each of the three file types, the files may differ in important ways with respect to the features. Those differences will be highlighted in subsequent sections of this tutorial.

Table 9-2 in the *SIPP Users' Guide* summarizes some of the file similarities and differences by topic.

### Changes in Variable Names

For the 1996 Panel, most variable names changed from those used in previous panels. When appropriate, the *SIPP Users' Guide* presents both sets of names.

The technical documentation that users receive with their data files will include an item booklet for the 1996 Panel and the paper survey instrument for earlier panels. **tip**

### The Survey Instrument and the Data Dictionary

With each order of a public use data file from the Census Bureau, users receive a set of technical documentation that includes, among other items, the survey instrument (or documentation of instrument screens and program code in the 1996 Panel) and a data dictionary.

**Survey Instrument.** The survey instrument is vital to understanding:

- What questions were asked
- How the questions were asked
- The order in which the questions were asked
- To whom the questions were asked
- The way in which the answers were recorded

## SIPP *tip*

Appendix A of the *SIPP Users' Guide* contains a crosswalk of variable names for the 1993 and 1996 core wave files. Link to a view of Appendix A.

**Section 1 - LABOR FORCE AND RECIPIENCY**

*(SHOW FLASHCARD J)*

1. During the 4-month period outlined on this calendar, that is, from 4 months ago through last month, did ... have a job or business, either full time or part time, even for only a few days? Mark "Yes" for active duty in the Armed Forces, any temporary or part-time work, and work without pay in a family business or farm.

1000 ☐ Yes - Mark "Worked" (code 170) on ISS and SKIP to 4  
 1001 ☐ No

2a. Even though ... did not have a job during this period, did ... spend any time looking for work or on layoff from a job?

1002 ☐ Yes  
 1003 ☐ No - SKIP to 3a

b. Please look at the calendar. In which weeks was ... looking for work or on layoff from a job? Please answer by giving the week number that appears to the right of each week on the calendar. Mark (X) all that apply.

1004	<input type="checkbox"/> ALL	1009	<input type="checkbox"/> 7	1010	<input type="checkbox"/> 13
1005	<input type="checkbox"/> 1	1010	<input type="checkbox"/> 8	1011	<input type="checkbox"/> 14
1006	<input type="checkbox"/> 2	1011	<input type="checkbox"/> 9	1012	<input type="checkbox"/> 15
1007	<input type="checkbox"/> 3	1012	<input type="checkbox"/> 10	1013	<input type="checkbox"/> 16
1008	<input type="checkbox"/> 4	1013	<input type="checkbox"/> 11	1014	<input type="checkbox"/> 17
1009	<input type="checkbox"/> 5	1014	<input type="checkbox"/> 12	1015	<input type="checkbox"/> 18
1010	<input type="checkbox"/> 6				

c. Could ... have taken a job during any of those weeks if one had been offered?

1042 ☐ Yes - SKIP to 3a  
 1043 ☐ No

d. What was the main reason ... could not take a job during those weeks? Mark (X) only one.

1044 ☐ Already had a job  
☐ Temporary illness  
☐ School  
☐ Other - Specify \_\_\_\_\_

3a. Even though ... did not have a job during this period, did ... do any work at all that earned some money?

1040 ☐ Yes - Mark "55" on ISS  
 1041 ☐ No - SKIP to 3a, page 4

b. In which of the months shown on this calendar did ... do that work? Mark (X) all that apply.

1048 ☐ Last month  
 1049 ☐ 2 months ago  
 1050 ☐ 3 months ago  
 1051 ☐ 4 months ago

4. Did ... have a job or business, either full or part time, during EACH of the weeks in this period? Note that the person did not have to work each week.

1052 ☐ Yes  
 1053 ☐ No - SKIP to 3a

5a. Was ... absent without pay from ...'s job or business for any FULL weeks during the 4-month period?

1054 ☐ Yes  
 1055 ☐ No - SKIP to 3a, page 4

b. Please look at the calendar. In which weeks was ... absent without pay? Please answer by giving the week number that appears to the right of each week on the calendar. Mark (X) all that apply.

1056	<input type="checkbox"/> ALL	1061	<input type="checkbox"/> 7	1062	<input type="checkbox"/> 13
1057	<input type="checkbox"/> 1	1062	<input type="checkbox"/> 8	1063	<input type="checkbox"/> 14
1058	<input type="checkbox"/> 2	1063	<input type="checkbox"/> 9	1064	<input type="checkbox"/> 15
1059	<input type="checkbox"/> 3	1064	<input type="checkbox"/> 10	1065	<input type="checkbox"/> 16
1060	<input type="checkbox"/> 4	1065	<input type="checkbox"/> 11	1066	<input type="checkbox"/> 17
1061	<input type="checkbox"/> 5	1066	<input type="checkbox"/> 12	1067	<input type="checkbox"/> 18
1062	<input type="checkbox"/> 6				

c. What was the main reason ... was absent without pay from ...'s job or business during those weeks? Mark (X) only one.

1068 ☐ On layoff  
☐ Own illness  
☐ On vacation  
☐ Bad weather  
☐ Labor dispute  
☐ New job to begin within 30 days  
☐ Other - Specify \_\_\_\_\_

NOTES

**Data Dictionary.** The data dictionary describes four aspects of each variable:

- Definition
- Sample universe for the corresponding survey question
- Ranges for all legal values
- Location in the file

It is important that users understand that the data dictionary does not replicate the survey instrument. Analysts should therefore be aware of the following:

- Variables on the data files do not have a one-to-one correspondence with questionnaire items.
- The range of possible values of variables on the data files does not always correspond exactly with the response categories in the survey instrument or the data dictionary.
- Variable names in the data dictionary may not readily reflect the variable's content.
- Skip patterns will not be obvious from simply looking at the data dictionary. *tip*

### Identification/Description Variables

#### Basic ID Variables in SIPP

The capacity to identify units across files allows SIPP users to:

- Follow participants over time
- Determine when an individual is present in the sample
- Verify the make-up of families and households

```

SURVEY OF INCOME AND PROGRAM PARTICIPATION,
1996 PANEL WAVE 1 TOPICAL MODULE DATA DICTIONARY

DATA      SIZE  BEGIN
D  SSUSEQ    5    1
T  SU: Sequence Number of Sample Unit - Primary
    Sort Key
U  All persons
V    1:50000 .Sequence Number

D  SSUID     12    6
T  SU: Sample Unit Identifier
    Sample Unit identifier This identifier is
    created by scrambling together the PSU,
    Segment, Serial, Serial Suffix of the
    original sample address. It may be used
    in matching sample units from different
    waves.
U  All persons
V  0000000000000:999999999999 .Scrambled Id

D  SPANEL    4    18
T  SU: Sample Code - Indicated Panel Year
U  All persons
V

```

### SIPP *tip*

*Analysts should become familiar with the survey instrument before using the data. This will prevent confusion and help avoid problems. It is also helpful to refer to the survey instrument and data dictionary while working with the data.*

The four most basic identification (ID) variables in SIPP include the following:

**Sample Unit IDs.** These uniquely identify each physical dwelling unit in the sample. The sample unit ID assigned to a person never changes. All people who have ever lived with a member of a given original sample unit share the same sample unit ID.

**Current Address IDs.** These identify the housing units occupied by one or more original sample members in a given month. They are assigned within sample units.

**Entry Address IDs.** These are the current address IDs for each sample member's initial address. They do not change when a person moves.


**Person Number IDs.** Person numbers are assigned sequentially, within each wave and each household, to all primary and secondary sample members when they first enter the sample.

These four variables have different names in the different types of public use files. [Link to a table that includes the names of the ID variables in the three types of files.](#)

### Monthly Interview Status

The monthly interview status variable, which has values of 0, 1, or 2, helps analysts determine whether or not to use the data for a person in a given month.

Analysts should use data only for those months in which a person's interview status is equal to 1. Examining either the weight variable or the variable used in the analysis itself, as is often done with other data sources, will lead the SIPP user astray. See Chapter 9 of the *SIPP Users' Guide* for more information.

Analysts should ignore any data for months in which a person's interview status is coded either 0 (indicating a person was not in the sample that month) or 2 (indicating a noninterview for that month). 

### SIPP *tip*

*Because the person-month core wave files and the 1996 topical module files contain records only for those months that a person has an interview status code of 1, the monthly interview status variables in those files can be safely ignored.*

## Identifying Persons

Analysts may need to identify which records belong to which individual in SIPP data files. For example, analysts may need that information to combine data from file types, to link family members, and to identify the recipient of government transfer income.

Each person in SIPP can be identified by the combination of sample unit ID, entry address ID, and person number. *tip*

## Identifying Households

A household consists of all people who occupy a housing unit, regardless of their relationships to one another. The many variations of households include, for example:

- A group of friends sharing a townhouse
- A single person in an apartment
- A family in a house

Each household contains one household reference person—the owner or renter of record.

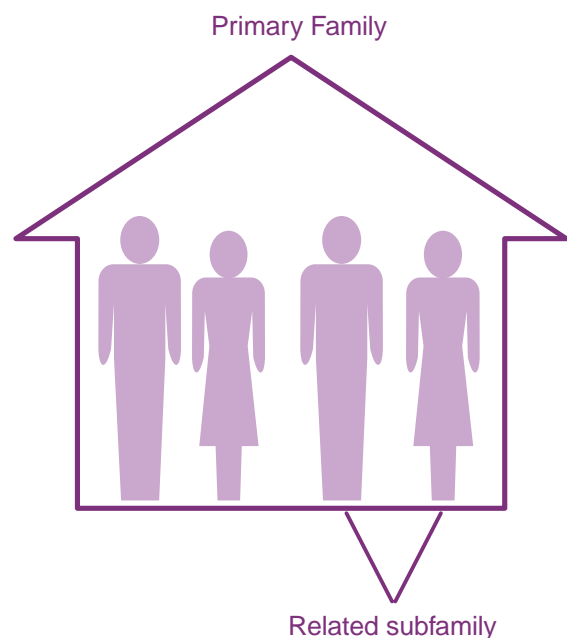
## Identifying Families

The Census Bureau defines a family as a group of two or more people who reside together and are related by birth, marriage, or adoption. There are several types of families that the Census Bureau distinguishes:

- A primary family contains the household reference person and all of his or her relatives.
- A related subfamily is a family unit within the primary family whose members are related to, but do not include, the household reference person. An example would be a son and his wife living with the son's parents, one of whom is the household reference person.

## SIPP *tip*

*For the 1996 Panel, analysts do not need to use the entry address to uniquely identify individuals.*



- An unrelated subfamily, or secondary family, is a family living in the household whose members are not related to the household reference person.
- A primary individual is a household reference person who lives alone or with nonrelatives. The Census Bureau sometimes treats primary individuals as one-person families and refers to them as pseudo-families.
- A secondary individual is not a household reference person and is not related to other people in the household. The Census Bureau also sometimes refers to such individuals as pseudo-families.

The Census Bureau has two principal methods for distinguishing families:

- The first method defines a family as all persons who are related and living together.
- The second method is similar to the first but excludes members of related subfamilies.

The variables and numbering schemes associated with these two methods allow analysts to construct various family units, including multigenerational families.

The various types of data files in SIPP, however, contain different identification information about family relationships. In fact, the topical module files contain no information for directly identifying different types of families. Thus, the analytic tasks for establishing family membership vary across file types. These differences will be highlighted in subsequent sections of the tutorial.

### **Describing Relationships to Reference Persons**

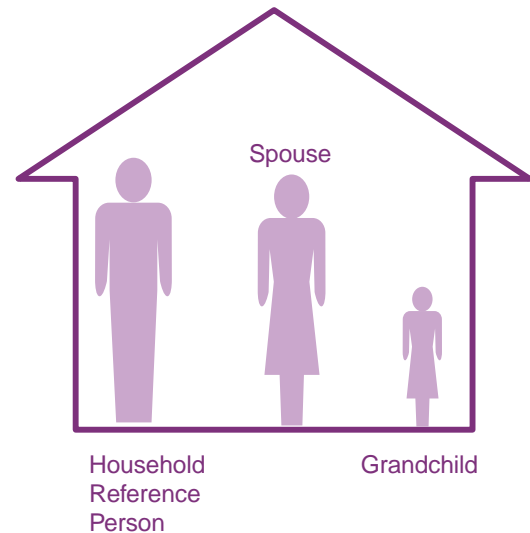
The SIPP data files contain variables that identify household and family reference persons. They also contain variables that describe how each person in the sample is related to the household reference person.



Users should note that the identity of the household reference person can change from one month to the next; thus, the household description could also change.

Analysts can use other relationship variables on the files to identify a variety of family configurations, such as households containing three generations.

The *SIPP Users' Guide* discusses important differences in the 1996 and pre-1996 relationship variables.



### Identifying Program Units

SIPP provides data for analyses involving program units for participants in transfer programs. SIPP records three characteristics regarding program participation:

- Whether the person is covered
- Who received the income or benefit
- The amount of the income or benefit

Coverage variables indicate whether a person is covered by a benefit directly or indirectly. For example, in a household receiving food stamps, the person who is the authorized recipient is identified as being covered directly.

Other members of the household are identified as being covered indirectly. Indirect recipients will have the same sample unit ID and current address ID as the primary recipient. **tip**

SIPP data also permit identification of members of common units within households, because most programs allow more than one program unit in a household. Members of common units can be identified by the sample unit ID and the authorized recipient variable. **tip**

Chapters 10–12 of the *SIPP Users' Guide* discuss specific variables related to program unit identification and exceptions to the rules for identifying program units.

### SIPP *tip*

*When a child receives a benefit, an adult will be the authorized recipient and will be flagged as not covered; the child will be flagged as covered. Except for WIC, no amounts of income or benefit are listed in the records of children under 15.*

### *tip*

*Unlike most transfer programs, Medicare is a person-based program in which each participant is an authorized recipient. Thus, SIPP files do not carry additional authorized recipient variables on the files.*



## Identifying Movers and Household Composition Changes

When SIPP original sample members move, sometimes changes in household composition occur. The mover may acquire a spouse, a roommate, a child, or other new household members. It may be important for analysts to know about these household composition changes during a particular reference period.

To identify movers, analysts should look for changes in current address fields. Except in rare cases (e.g., merged households), movers' other basic ID variables—sample unit ID, entry address ID, and person number—remain the same. **tip**

Chapters 10–12 of the *SIPP Users' Guide* contain tables and explanatory text that illustrate how analysts can identify and track movers.

## Identifying States and Metropolitan Areas

**States.** Even though it is possible to identify most states, SIPP was not designed to be representative at the state level. Therefore, SIPP data should not be used to produce state-level estimates.

**Metropolitan Areas.** Analysts can use variables in the core wave files to produce national estimates of the metropolitan population and to identify 93 Metropolitan Statistical Areas and Consolidated Metropolitan Statistical Areas.

**Nonmetropolitan Areas.** The Census Bureau recodes a small random sample of metropolitan households as nonmetropolitan households to protect respondent confidentiality. Thus, SIPP data cannot be used to produce national estimates of the nonmetropolitan population.

## SIPP *tip*

*In the pre-1996 panels, when two SIPP households merged, or when one split but then recombined with new secondary sample members, some sample members may have received new ID variables. Because of the rarity of these cases, the 1996 Panel files do not include information about them.*

## Choosing Weights

SIPP samples different households and people at different rates. Consequently, analysts should use weights to reduce the likelihood of biased estimates of population characteristics.

SIPP data files include a number of alternative weights. The choice of the appropriate weight for an analysis depends on the population of interest—person, household, family, and so on.

Analysts need to ask:

1. Which sample or subsample of SIPP is the basis for the estimate?
2. What population does the sample represent?

To obtain weights, analysts should check the files they are using:

- Weights for each calendar month covered by a panel are in the core wave files.
- A single weight appears in the topical module files. **tip**
- Weights for calendar years are on the longitudinal files.

The source and accuracy statements that accompany the three types of files include suggestions about which weights to use and how to use them, as does Chapter 8 of the *SIPP Users' Guide*.

## Income Topcoding

To protect the confidentiality of SIPP respondents, the Census Bureau topcodes very high incomes on the public use data files. New income topcoding procedures were instituted with the 1996 Panel.



## SIPP *tip*

*Before 1996, the weight on the topical module files is the person interview month weight for those who provided data for the module. In the 1996 Panel, the weight on the topical module file is the person cross-sectional weight for the fourth reference month.*

## 1996 Panel

**Unearned Income.** When the total amount of asset income or of certain types of general income for a wave exceeds the established ceiling, the monthly amounts in excess of the monthly threshold are replaced by monthly topcode values. *tip*

**Employment Income.** Monthly employment income falls into three categories within SIPP:

- Wage and salary income
- Self-employed earnings
- Other worker arrangements

Each of these three sources was topcoded separately.

In the 1996 Panel, the method used to topcode employment income is based on the mean of reported unweighted amounts above the threshold in Wave 1 of the panel.

An algorithm was used to establish topcode values for 12 cells of different combinations of gender, race, and employment status. Each respondent's topcode value is assigned in accordance with his or her corresponding cell. *tip*

The topcode amounts established in Wave 1 of the 1996 Panel were used for all waves of the panel, with a wave adjustment, determined by formula, for inflation and real growth in earned income.

## Pre-1996 Panels

In earlier panels, the topcode amount for the wave was \$33,332; thus, in most cases, the topcode amount for monthly income was \$8,333.

Income from various sources (multiple jobs, businesses, property) was not independently topcoded in the pre-1996 panels.

## SIPP *tip*

*Not all income sources are topcoded. For example, the amount of food stamp income is not topcoded. See Appendix B of the SIPP Users' Guide for a list of topcoded income variables in the 1996 Panel.*

## *tip*

*Chapter 10 of the SIPP Users' Guide contains a discussion of the 1996 income topcoding method and examples illustrating its application.*

## **Using Allocation Flags**

As discussed earlier in the tutorial, the Census Bureau often imputes information when a person does not respond to the survey or to a particular question.

When a variable is imputed, the Census Bureau sets an allocation, or imputation, flag to identify the imputed variable. Variables selected for imputation vary across the three types of files.

Not all imputations are readily apparent, however.

**Whole Record Imputation.** Whole records were sometimes imputed with the Type Z procedure when person-level interviews were not successfully conducted. The variables needed to identify these records vary across the file types.

**EPPFLAG and Little Type Z Imputation.** In the 1996 Panel, the Census Bureau used special imputation procedures, known as EPPFLAG and little Type Z, for labor force items. The allocation flags for items imputed with these procedures will not indicate by themselves the imputation status of the items.

Analysts should read the discussion on allocation flags in Chapter 4 of the *SIPP Users' Guide* to learn how to identify items imputed with these special procedures.

**Composite Variables.** Variables are imputed and the allocation (imputation) flags are set before the creation of composite variables, such as household and family aggregates. Since total household income is computed after person-level imputation has occurred, total household income may be based, in part, on imputed information. There will be no direct indication, though, on the records of other household members that any information on household income has been imputed.

Analysts should use the person-level imputation flags of all household and family members to identify aggregate amounts that include imputed values.